Cahier des Charges

pour

une prestation de conception de chaîne de traitement, de bases de données et de site Web

Unité de Recherche : DIA-PC

2007

Christine Dubreuil-Tranchant

Institut de Recherche pour le Développement - IRD

911, avenue Agropolis

34000 Montpellier Cedex – France

Chapitre 1. INFORMATIONS GENERALES

1. Présentation de l'Institut de Recherche pour le Développement

L'Institut de Recherche pour le Développement (IRD) est un Etablissement Public à caractère Scientifique et Technologique (EPST), placé sous la tutelle des ministères de la Recherche et des Affaires Etrangères. Depuis soixante ans, l'Institut conduit des recherches sur les milieux intertropicaux qui sont devenues des références internationales. L'IRD mène des recherches en Afrique, dans l'Océan Indien, en Amérique latine et dans le Pacifique.

Il conduit des missions de recherche sur :

- L'environnement et grands écosystèmes
- Agriculture en milieux tropicaux fragiles
- Environnement et santé
- Hommes et sociétés en mutation

Près de 2600 personnes travaillent pour l'Institut, dont 45% en France Métropolitaine (les autres personnels étant répartis dans les DOM-TOM et 38 pays étrangers). Il dispose d'implantations dans 26 pays de la zone intertropicale. Il compte également cinq implantations en métropole et cinq dans les DOM-TOM.

2. Présentation du projet de l'unité de recherche « DIA-PC »

Depuis plusieurs années, les équipes du domaine végétal de l'IRD appartenant à différentes UMRs/URs développent des projets de génomique concernant des plantes tropicales d'intérêts majeurs pour les pays du SUD (Coffea, Hevea, Arécacées, Casuarinacées, ...) et s'articulent essentiellement autour de 2 plantes modèles entièrement séquencées Arabidopsis thaliana et Oryza sativa.
La production en masse de données génomiques et protéomiques et la nécessité de leur analyse a conduit à la mise en place d'une plate-forme bio-informatique dédiée à la génomique végétale sur le centre IRD de Montpellier (Resp. C. Tranchant-Dubreuil).

Cette plate-forme centralise de nombreuses ressources bio-informatiques :

- des logiciels classiques de bio-informatique (alignements de séquences, phylogénie etc.)
- des banques de séquences publiques et privées
- des programmes et des systèmes d'information développés par le service bio-informatique, au sein de nos UMR, permettant respectivement de traiter des volumes importants de données brutes issues des expérimentations et de mieux gérer/exploiter ces masses importantes de données.

Depuis 2001, de nombreux projets de séquençage d'ESTs ont été réalisés par les équipes végétales et une chaîne de traitement ou pipeline d'annotation d'ESTs utilisant différents logiciels bio-informatiques gratuits (Blast, Stackpack...) a été conçue pour analyser ces volumes importants d'informations. Les données brutes placées en entrée du pipeline ainsi que les données produites par le pipeline sont stockées dans une base de données (EST-db), ce qui permet de gérer et d'exploiter les données produites sur les différentes plantes tropicales. Les données sont accessibles via un site web associé à la base de données. Différentes équipes ont déjà valoriser l'utilisation du pipeline au travers de publications scientifiques (Voir Paragraphe « Liste des Publications »). Cet outil est en constante évolution et c'est dans ce contexte que s'inscrit ce projet qui propose une optimisation de l'outil existant couplé à l'ajout de nouvelles fonctions.

Plusieurs équipes se sont maintenant engagées dans des projets de génomique comparative et la réalisation de ces programmes nécessite le développement d'un nouveau module au pipeline dédié à la réalisation d'analyses de génomique comparative inter et intra-espèces. Ces projets concernent en priorité les différentes espèces propres à l'IRD et s'étendront à des espèces travaillées dans des équipes d'autres organismes (INRA, CNRS, ...) associées à l'IRD par le biais des UMRs ou de projet de recherche communs. Une brève description des programmes de recherche concernés est proposée (Voir paragraphe « Descriptif des Projets Scientifiques») et permettra de mieux mesurer la nécessité de développer cette nouvelle fonctionnalité.
Cependant, le développement de ce nouveau module est conditionné par l'amélioration préalable du pipeline existant afin, d'une part, de le rendre plus convivial, totalement générique et, d'autre part, d'y ajouter certaines fonctionnalités, telle que l'annotation automatique des séquences en grandes fonctions selon Gene Ontology ou encore un affichage graphique des séquences des ESTs et de leur contigue. Enfin, pour valoriser l'outil EST-db, une étape de documentation est indispensable pour sa déclaration à l'Agence de Protection des Logiciels en vue de sa distribution.

L'objectif de ce projet est de réaliser ces différentes opérations afin d'obtenir un outil EST-db convivial, générique, facilement transférable à d'autres équipes (IRD ou extérieure) et répondant aux demandes actuelles des projets scientifiques (traitements de données de séquençage, analyses comparatives) en y ajoutant des nouvelles fonctionnalités).

Chapitre 2. DESCRITION DE L'APPLICATION ESTdb

1. Description d'ESTdb

En 2001, suite à la réalisation de plusieurs projets de séquençage d'ESTs, le développement d'ESTdb a été initié et cette application se compose, d'une part, d'une chaîne de traitement permettant d'analyser les ESTs et, d'autre part, d'une base de données (et d'un site web associé) destiné à gérer et mieux exploiter les informations générées par la chaîne de traitement. L'application a été développée au cours de 5 stages de master « Informatique Pour les Sciences » de l'Université de Montpellier.
Elle est installé sur la plate-forme bio-informatique de l'IRD et est utilisée par les 4 UMRs du domaine végétal basées sur le centre IRD de Montpellier ainsi que par des partenaires et IRDiens expatriés (ex : Projet Hévéa - Thaïlande, Projet Café - Ile de la Réunion). Le pipeline a aussi été utilisé pour analyser des données d'autres UMRs telles que des ESTs issues de la souris. A l'heure actuelle, plus de 20000 ESTs ont été analysées et 200000 données associées à ces séquences ont été générées. Ce volume de données ne cesse d'augmenter et de nouvelles analyses sont demandées par les chercheurs. Il s'avère donc nécessaire de développer de nouvelles fonctionnalités sur l'outil EST-db.

2. Description technique du pipeline

Ce programme écrit en perl/bioperl permet de combiner l'exécution de plusieurs logiciels, l'analyse des résultats générés et de réaliser d'autres fonctionnalités répondant à des critères propres au laboratoire. Les données brutes et les données générées sont ensuite stockées dans une base de données MySQL. Compte tenu du volume important de données à analyser et de traitements, ce pipeline automatise l'analyse de chaque EST :

- A l'issue du séquençage, les chromatogrammes des ESTs sont analysés afin d'obtenir la séquence nucléique à l'aide du logiciel de « base calling » Phred.
- La séquence est ensuite analysée afin de masquer les bases de mauvaise qualité (les fichiers phd.1 sont analysés et les bases ayant une qualité inférieure à une valeur seuil paramétrée sont remplacées par des N). Les séquences appartenant au vecteur sont ensuite détectées à l'aide du logiciel Vecscreen puis elles sont masquées et supprimées. Les queues polyA sont également supprimées. Les extrémités de séquences de mauvaise qualité (seuil paramétrable lors du lancement du pipeline) sont également supprimées et les séquences de petite taille (seuil parémétrable) sont éliminées.
- Afin de supprimer la redondance au niveau des séquences, une phase de contiguage est nécessaire. Ceci est réalisé par le logiciel Stackpack. A l'issue du contigage, les ESTs appartiennent ou non à un contig.
- Puis, l'étape suivante est l'annotation des séquences qui doit renseigner sur la fonction des protéines putatives éventuellement associées : la séquence d'EST est comparée à une banque de séquences (Blast).

Les résultats de chaque étape de ce traitement sont archivés dans une base de donnée et l'ensemble (pipeline – base de donnée) est consultable et utilisable au travers d'une interface Web.

3. Liste des logiciels

Phred : Le programme Phred, développé en C, est un logiciel de « base calling » qui se base sur la méthode de Fournier pour lire les 4 courbes du chromatogramme. Il appelle une à une les bases, leur assigne une valeur de qualité et écrit les résultats dans des fichiers de sortie. http://www.phrap.org/phredphrapconsed.html

Vecscreen : Cet outil disponible sur NCBI permet d'identifier rapidement des segments de séquences nucléiques d'origine vectorielle. Il recherche la position du vecteur dans une séquence en utilisant le programme Blast automatiquement paramétré pour une détection optimale des contaminations. http://www.ncbi.nlm.nih.gov/VecScreen/

Stackpack : Ce logiciel possède un programme réalisant le clustering, l'assemblage de séquences présentant de courtes régions chevauchantes. Il regroupe 3 algorithmes :

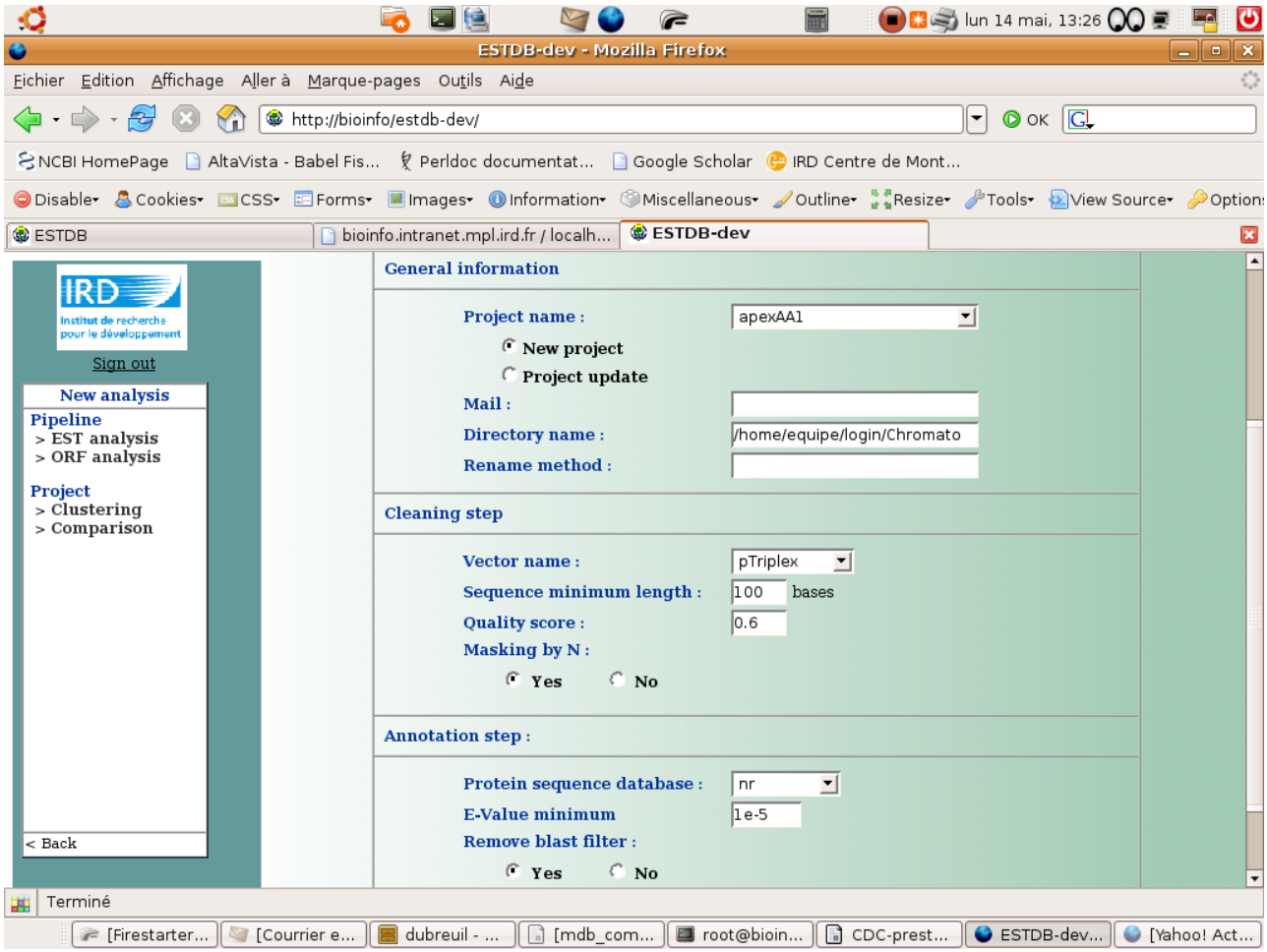
- l'algorithme agglomératif D2_cluster, plus rapide que blast est utilisé pour le clutering initial; les séquences doivent être longues et seules les grandes similarités sont détectées.
- l'algorithme du programme phrap aligne rapidement toutes les séquences d'un cluster entre elles mais les informations sur la variation à l'intérieur même du cluster sont insuffisantes pour établir une séquence consensus.
- le programme craw intervient dans l'ultime étape pour analyser l'alignement et déterminer la séquence consensus. Ce logiciel est développé en python et toutes les informations manipulées par ce logiciel sont stockées dans une base de données Mysql. http://www.sanbi.ac.za/Dbases.html

BLAST (Basic Local Alignment Search Tool) est un programme de recherche de similarité développé au NCBI/ Genbank. http://www.ncbi.nlm.nih.gov/Tools/

4. Description de l'application web

L'application web permet :

- de gérer les différents projets d'ESTs qui sont accessibles à un groupe d'utilisateurs identifiés avec des droits définis. Les profils sont les suivants :
  - super-administrateur : l'utilisateur peut créér/modifier/effacer des projets/ - ajouter ou modifier un nouvel utilisateur et son profil, gérer les droits d'accès d'un utilisateur à des projets - lancer ou mettre à jour une analyse en lançant le pipeline d'ESTs – Consulter les données de tous les projets
  - administrateur : l'utilisateur peut lancer ou mettre à jour une analyse en lançant le pipeline d'ESTs – Consulter les données des projets auquels il a accès
  - utilisateur : il peut consulter les données des projets auquels il a accès.
- de lancer le pipeline d'ESTs au travers d'un formulaire en précisant les paramètres suivants :



- de consulter les informations relatives à un projet notamment :
  1. les statistiques (nombre d'ESTs traitées, valides, nombres d'unigènes (singleton et contigs)...
  2. un tableau résumant pour l'ensemble des ESTs et des contigues les annotations blast permettant d'accéder soit à la fiche blast soit à la fiche descriptive de la séquence
  3. Exporter les séquences d'ESTs et de contigs au format fasta
  4. Rechercher une séquence par son nom
  5. Rechercher les séquences en tapant un mot clé et permettant de lister les séquences dont l'annotation blast contient ce mot clé dans plusieurs projets

6. Blaster une séquence contre des banques d'ESTs

5. Etat des lieux

L'ensemble du code a été repris progressivement et réorganisés selon l'arborescence suivante :

- répertoire web : contient toutes les interfaces web de l'application ainsi que le script perl permettant de lancer le pipeline. Au fur et à mesure de la reprise du code, le code a été réorganisé, structuré et commenté de manière succincte en français.
- répertoire biolibrairies : contient différents fichiers et objets perl utilisés par le pipeline et l'application web
  - un fichier centralisant les requêtes SQL
  - un fichier centralisant plusieurs fonctions perl propres à ESTdb
  - un fichier centraiisant des variables telles que les paramètres de connexion à la base de données ESTdb, les temps de sessions, les noms de répertoires au niveau desquels sont stockées les données intermédiaires générées par le pipeline par exemple.
  - les différents objets perl utilisés par le pipeline
- répertoire tmp/estdb au niveau duquel sont stockés les données temporaires

Un des objectifs est :

- de poursuivre ce travail de documentation des programmes (commentaires à ajouter ou traduire en anglais), de rédiger un manuel d'installation en anglais et une documentation technique.
- de bien veiller à la généricité du code et à une installation facile et souple de la base de données, de l'interface web et du pipeline.

Cet objectif doit être réalisé en vue de la déclaration de l'application ESTdb auprès de l'Agence de Protection des Logiciels afin de pouvoir distribuer ESTdb aux partenaires et à la communauté scientifique tout en veillant au respect des droits d'auteurs et de la propriété intellectuelle.

Chapitre 3. DEVELOPPEMENT A REALISER AU NIVEAU d'ESTdb

1. Description rapide des livrables souhaités pour l'application ESTdb

Le projet SPIRALE que nous proposons a pour objectifs :

- de finaliser le développement de l'application ESTdb (documentation du code (chaîne de traitement et objets perl associés, interfaces web associées à la base de données), code suffisament générique/paramétrable pour que l'application soit facilement transférable sur une autre plate-forme bio-informatique, ajout de nouvelles fonctions qui donneront une plus value importante à l'application) en vue de la déclarer à l'Agence de Protection des Logiciels et de la distribuer aux autres URs de l'IRD et partenaires intéressés par l'outil
- d'améliorer la version actuelle. Une phase d'analyse des besoins devra être menée au début du projet pour déterminer exactement les améliorations à apporter à l'application actuelle. Voici les fonctionnalités demandées par plusieurs utilsateurs :
  - Lancement du pipeline à partir des séquences (format fasta) et pas uniquement des chromatogrammes.
  - Amélioration des interfaces de consultation : ceci concerne particulièrement la visualisation des données issues du traitement par le logiciel gratuit Stackpack. En effet, l'installation de la dernière version de ce logiciel a eu de nombreuses conséquences dont la modification des fichiers de résultats générés par Stackpack. Les résultats affichés actuellement sont insuffisants et une adaptation du système est nécessaire. Devra être ajouté une interface graphique permettant de visualiser des séquences ESTs comparées à leur contig.
  - Amélioration de l'interface d'administration. A l'heure actuelle, seul l'administrateur (le bio-informaticien) peut créer les projets et les comptes utilisateurs. Il faudrait mettre en place un acteur gestionnaire de données qui puissent créer et modifier uniquement ses projets. Il pourrait également gérer les comptes utilisateur au niveau de ses projets.
- de développer deux nouveaux modules d'analyse à ESTdb dédiés à l'annotation automatique des séquences et à la réalisation d'analyse de génomique comparative :
  - le module "Annotation Automatique" doit permettre de classer automatiquement les séquences ESTs annotées par le logiciel Blast en grande fonction en se basant sur une ontologie du domaine, la Gene Ontology. Une visualisation graphique de la répartition des ESTs en grande fonction est demandée.
  - le module "Génomique comparative" permettra de réaliser des comparaisons entre les génomes des diverses plantes étudiées et/ou avec les génomes des plantes modèles. Ces analyses optimiseront l'identification de séquences orthologues entre différentes espèces et donc l'annotation des gènes identifiés. Ces comparaisons permettront de rechercher les relations existantes entre les gènes de différentes espèces (synténie) ainsi que les relations de ces gènes au sein d'un même génome.

2. Description détaillée des livrables dans l'ordre de priorité souhaité (à affiner)

Livrable 1 : Finaliser le développement de l'application ESTdb (documentation du code (chaîne de traitement et objets perl associés, interfaces web associées à la base de données), code suffisament générique/paramétrable pour que l'application soit facilement transférable sur une autre plate-forme bio-informatique, ajout de nouvelles fonctions qui donneront une plus value importante à l'application) en vue de la déclarer à l'Agence de Protection des Logiciels et de la distribuer aux autres URs de l'IRD et partenaires intéressés par l'outil.

A réaliser :

- compléter la documentation du code actuellement succincte et réalisée en français. Tout ce travail devra être fait en anglais
- rédiger la documentation technique et le manuel d'installation de l'application en anglais
- actualiser le code afin que la base de données, l'application web et le pipeline soient facilement installable et configurable. Ce travail a déjà été bien commencé.

Livrable 2 : Améliorer la version actuelle. Une phase d'analyse des besoins devra être menée au début du projet pour déterminer exactement les améliorations à apporter à l'application actuelle. Pour réaliser cette phase, une réunion sera organisée avec 5-6 personnes, actuellement déjà membre du comité bio-informatique et réprésentant leur équipe au niveau du comité (équipes impliquées pour participer au groupe de travail: Valérie Hocher-Florence Auguy/équipe « casuarina », Tim Transbarger/équipe « palmier », Diana Fernadez/équipe « résistance », Valérie Poncet équipe « café ». Ce groupe de travail sera coordonnée par C. Dubreuil (coté bio-informatique) et valérie Hocher (côté biologie). A ce groupe de travail, participera la DSI et la société ASA. Il sera demandé aux membres de ce groupe de communiquer leurs besoins aux coordinatrices et

les besoins/fonctionnalités attendues seront discutées et validée ou non lors d'une réunion du groupe de travail.

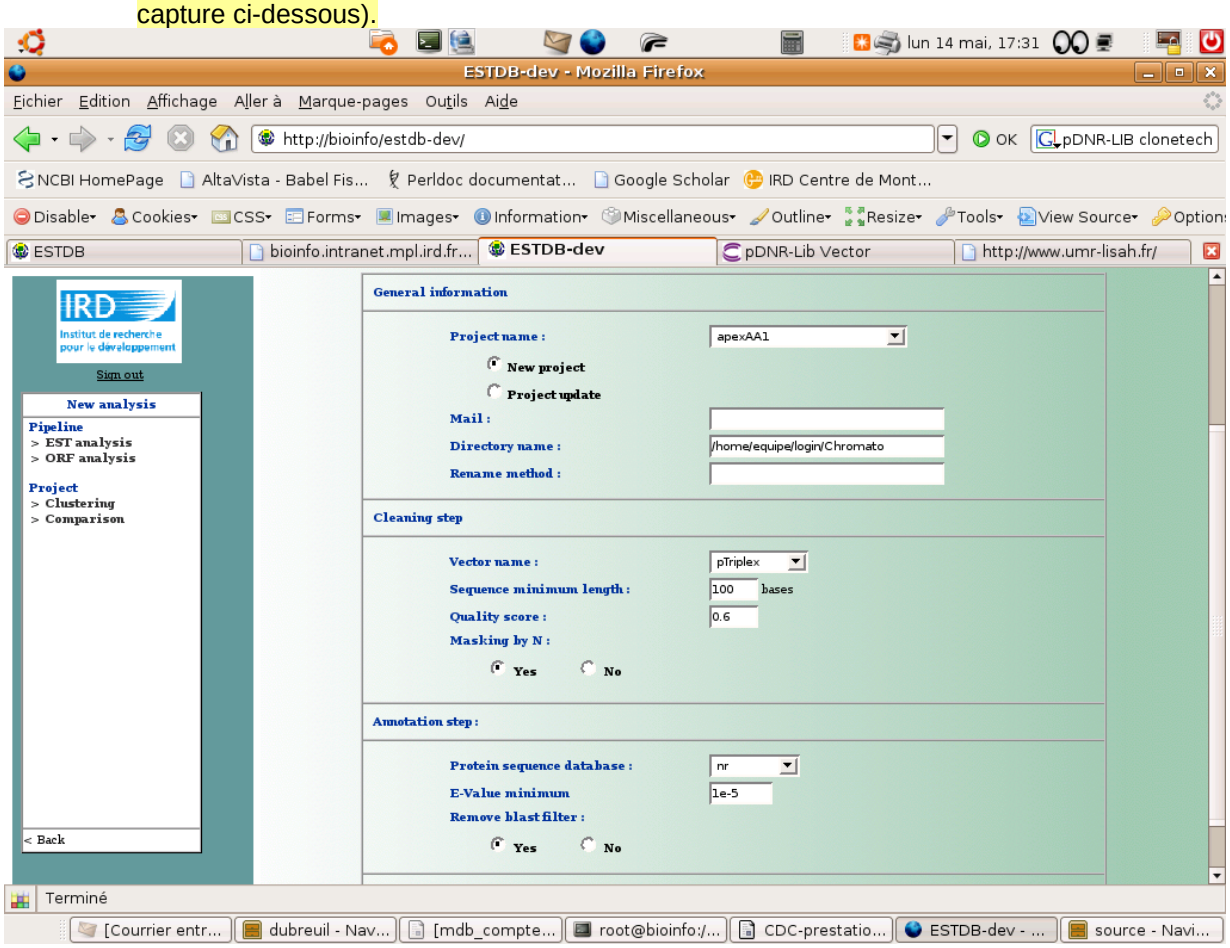
D'ores et déjà, voici les fonctionnalités déjà demandées par plusieurs utilisateurs (mais ces fonctionnalités devront être validées par le groupe de travail) :

- Amélioration des interfaces de lancement du pipeline d'ESTs et/ou du pipeline :

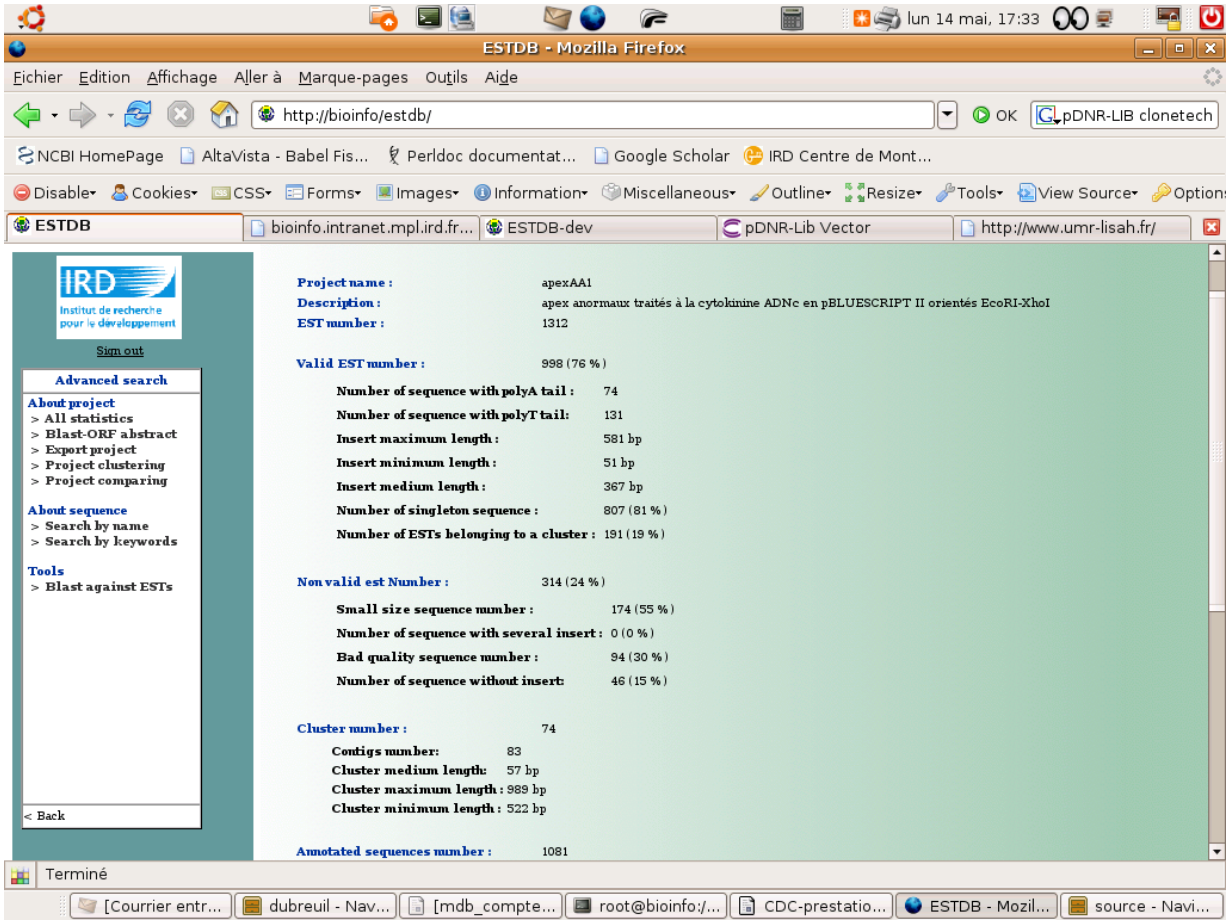
  1) Lancement du pipeline à partir des séquences (format fasta) et pas uniquement des chromatogrammes. Au lieu de partir de l'étape des chromatogrammes, les utilisateurs pourront mettre en entrée un fichier au format fasta contenant toutes leurs séquences d'ESTs et directement passer à l'étape Vecscreen. Avec cette fonctionnalité, aucune information est générée par l'étape de phred et il faudra adapter les programmes au niveau de l'insertion des données dans la base de données (les données relatives à l'étape de phred seront absentes)

  2) Actuellement, seul un blastx est fait contre nr lors de l'anaylse blast du pipeline. Si le blastx d'une séquence contre nr donne aucun résultat, aucune autre analyse n'est faite. Les utilisateurs souhaiteraient que si cette séquence ne donne pas de résultat avec un blastx, un blastn contre nt (ou autre base de données?) soit fait pour cette séquence.

  3) Modification de la base de données pour sauvegarder les paramètres de lancement du pipeline qui sont indiquées dans l'interface de lancement du pipeline pour un projet d'ESTs (cf. capture ci-dessous).



Ces paramètres devront apparaître au niveau de la page statistique d'un projet (cf. capture ci dessous).

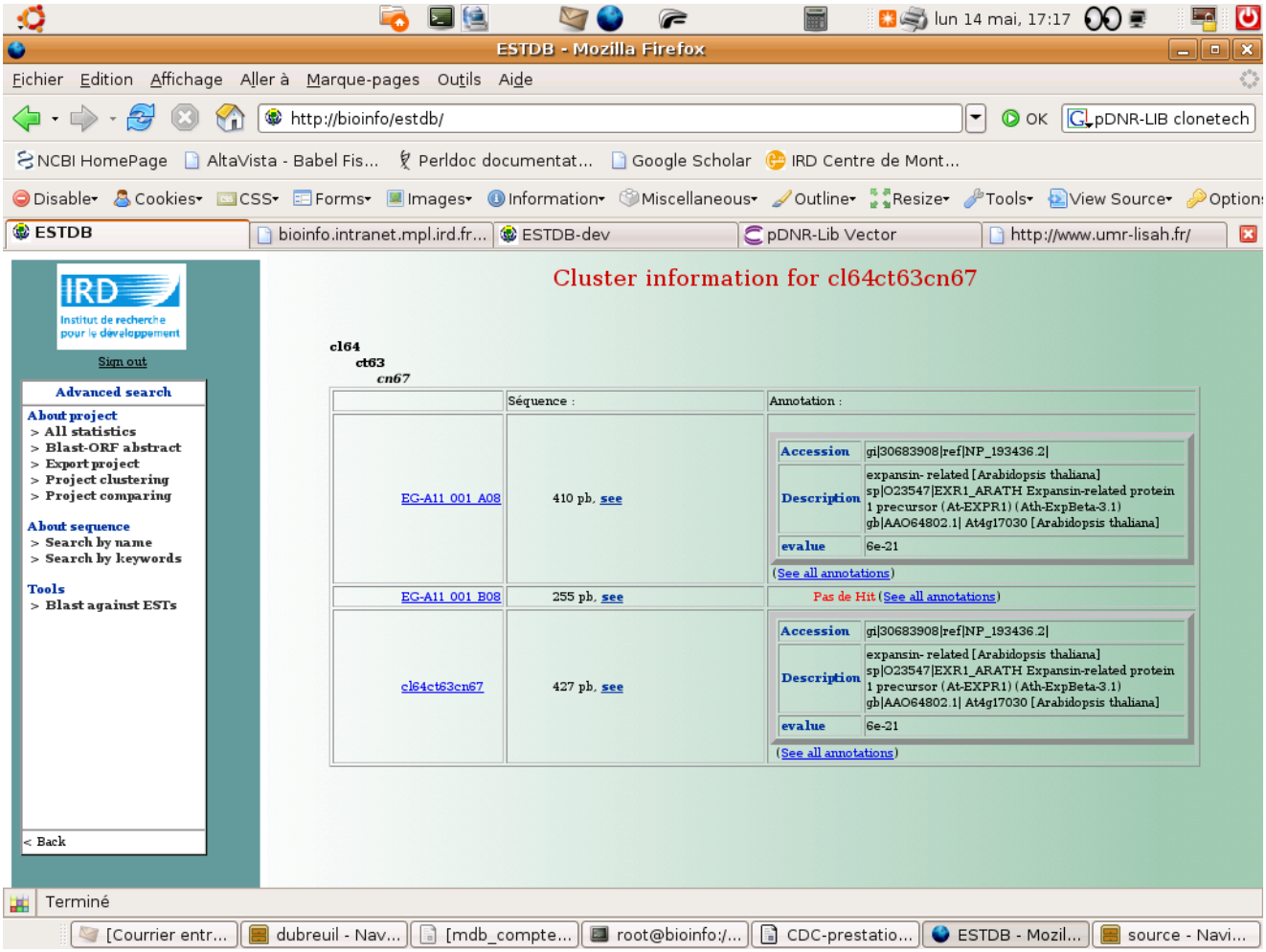

ESTDB - Mozilla Firefox
http://bioinfo/estdb/
ESTDB
Sign out
Advanced search
About project
> All statistics
> Blast-ORF abstract
> Export project
> Project clustering
> Project comparing
About sequence
> Search by name
> Search by keywords
Tools
> Blast against ESTs
< Back
Project name : apexAA1
Description : apex anormaux traités à la cytokinine ADNc en pBLUESCRIPT II orientés EcoRI-XhoI
EST number : 1312
Valid EST number : 998 (76 %)
Number of sequence with polyA tail : 74
Number of sequence with polyT tail: 131
Insert maximum length : 581 bp
Insert minimum length : 51 bp
Insert medium length : 367 bp
Number of singleton sequence : 807 (81 %)
Number of ESTs belonging to a cluster : 191 (19 %)
Non valid est Number : 314 (24 %)
Small size sequence number : 174 (55 %)
Number of sequence with several insert : 0 (0 %)
Bad quality sequence number : 94 (30 %)
Number of sequence without insert: 46 (15 %)
Cluster number : 74
Contigs number: 83
Cluster medium length: 57 bp
Cluster maximum length : 989 bp
Cluster minimum length : 522 bp
Annotated sequences number : 1081
Terminé

4) Trouver une solution pour le problème d'identifiant des contigues (attribués automatiquement par stackpack) et qui changent à chaque mise à jour du projet avec perte des fonctions et grandes fonctions déjà attribuées aux contigues (par exemple, lancement du pipeline de nouveau car une plaque d'ESTs a été séquencée et les chromatogrammes doivent être ajoutés au projet). Si le nouveau contigue contient majoritairement les mêmes ESTs qu'un contigue avant l'analyse ainsi que la même description blast, on doit pouvoir conserver l'annotation en grande fonction et avoir une trace de l'ancien nom du contigue. Fonctionnalité à discuter en groupe de travail pour les modalités et la priorité.

5) Mettre en place une procédure qui relance automatiquement le blast des séquences d'ESTs au niveau des différents projets 2 fois par an (à discuter de la fréquence de la mise à jour avec le groupe de travail) avec un système d'alarme pour les nouvelles annotations.

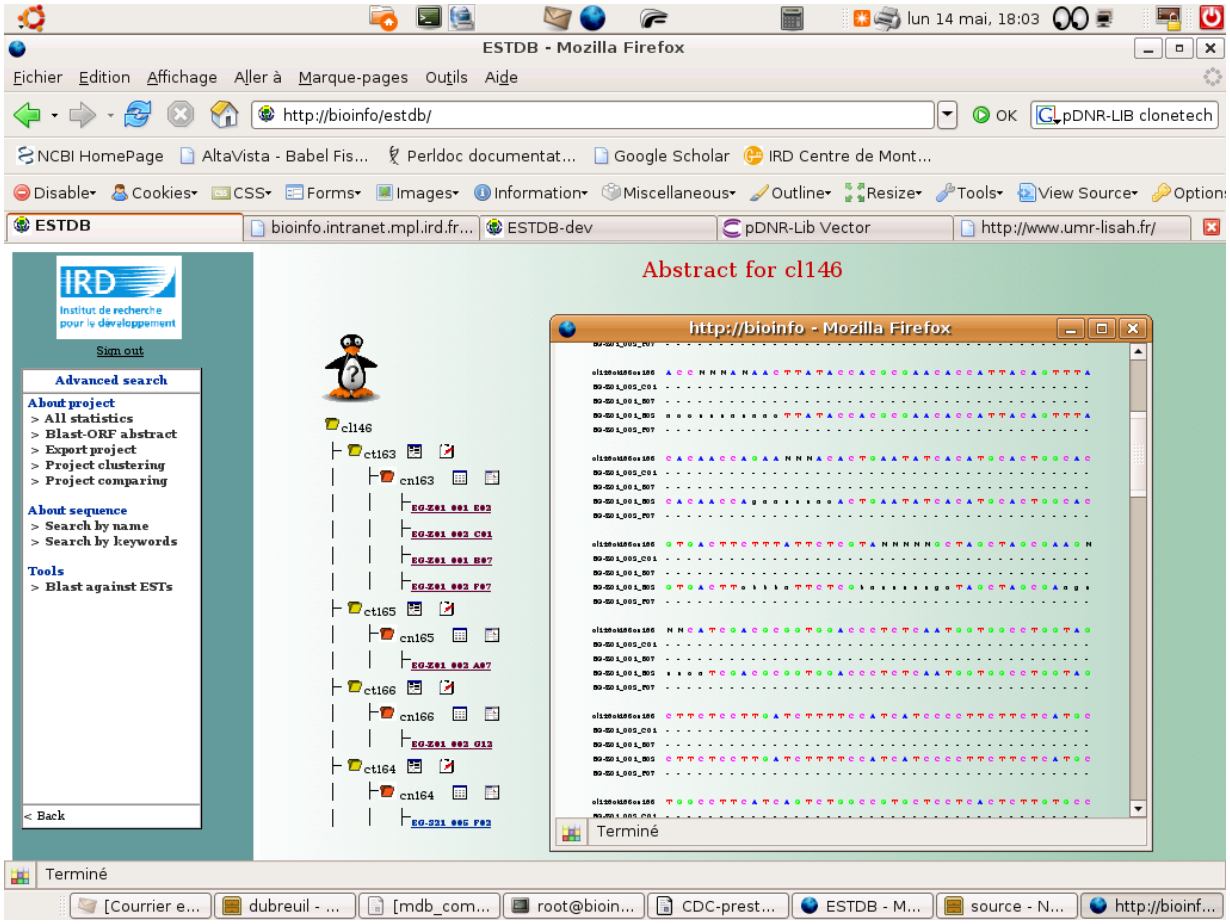
– Amélioration des interfaces de consultation :

1) Interface « cluster information » : ceci concerne particulièrement la visualisation des données issues du traitement par le logiciel gratuit Stackpack. En effet, l'installation de la dernière version de ce logiciel a eu de nombreuses conséquences dont la modification des fichiers de résultats générés par Stackpack. Avec l'ancienne version, le pipeline récupérait directement l'alignement de la séquence consensus avec les ESTs. Or la nouvelle version de stackpack (et donc de la base de données associées à stackpack) a évolué et actuellement cette information n'est plus récupérée. L'objectif est de récupérer cette information de nouveau (alignement directement comme précédemment? position des ESTs sur la séquence consensus et affichage graphique en récupérant ces informations dans la base de données ESTDb?). Les résultats affichés actuellement sont insuffisants et une adaptation du système est nécessaire.



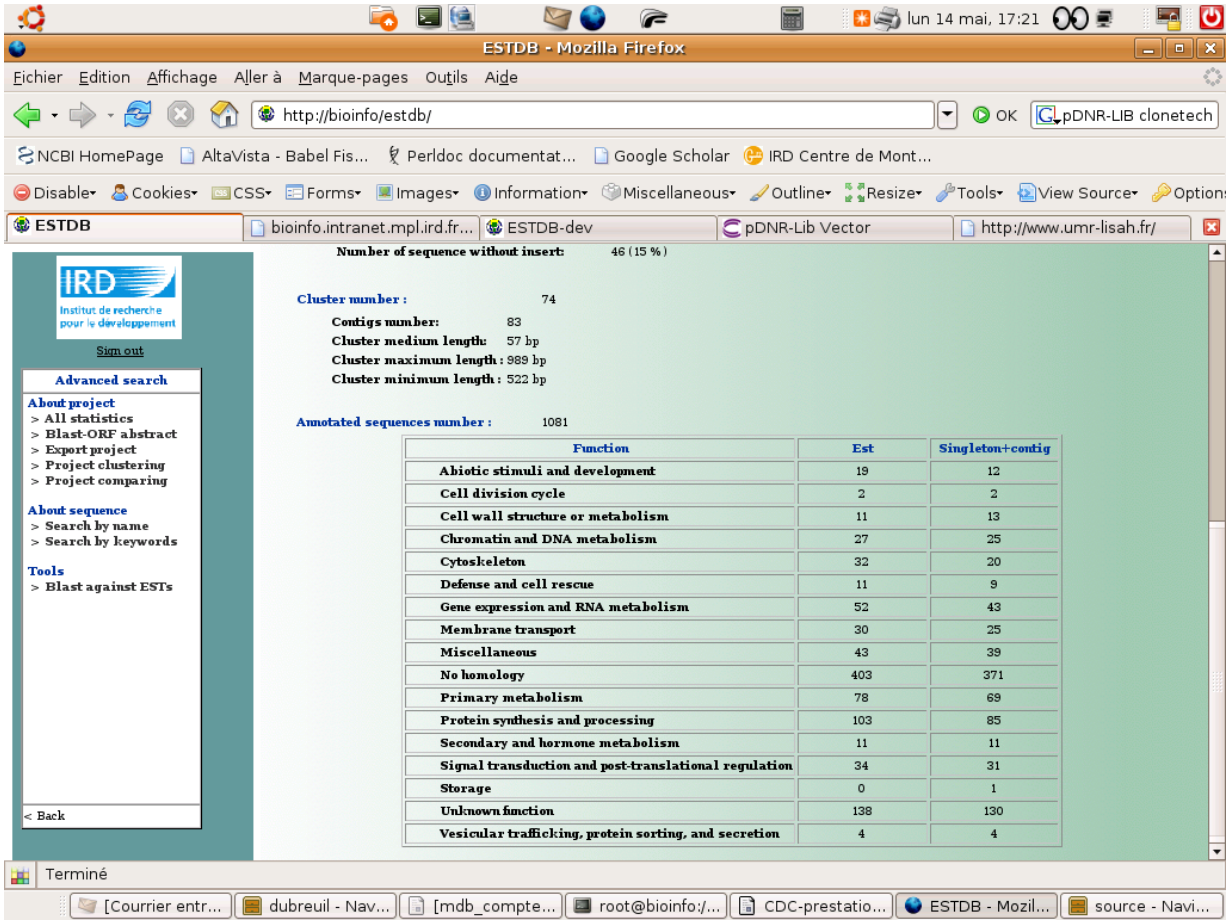
capture d'écran de la page « cluster information » : il faudrait pouvoir visualiser la séquence consensus et l'alignement des 2 ESTs vis à vis de cette séquence consensus.

Les interfaces développées avec la version précédente de stackpack permettait d'avoir ce type de page :



Ce type de page est souhaité que ce soit pour visualiser les informations de clustering d'ESTs au sein d'un même projet d'ESTs ou au sein de plusieurs projets d'ESTs.

– 2) Interface « statistique » : cette page donne différentes informations « statistiques » pour un projet donné (nombre d'ESTs traitées, valides, non valides, nombre d'unigènes etc.). Il donne aussi un tableau résumant le classement en grande fonction (réalisé à l'heure actuelle manuellement) comme le montre la capture d'écran suivante :



Function	Est	Singleton+contig
Abiotic stimuli and development	19	12
Cell division cycle	2	2
Cell wall structure or metabolism	11	13
Chromatin and DNA metabolism	27	25
Cytoskeleton	32	20
Defense and cell rescue	11	9
Gene expression and RNA metabolism	52	43
Membrane transport	30	25
Miscellaneous	43	39
No homology	403	371
Primary metabolism	78	69
Protein synthesis and processing	103	85
Secondary and hormone metabolism	11	11
Signal transduction and post-translational regulation	34	31
Storage	0	1
Unknown function	138	130
Vesicular trafficking, protein sorting, and secretion	4	4

Il est souhaité de pouvoir obtenir pour une grande fonction donnée :

– le nombre d'ESTs classées dans cette grande fonction (appartenant ou non à un contigue)
– le nombre d'ESTs singleton classées dans cette grande fonction
– le nombre de contigues classés dans cette grande fonction

Il est souhaité aussi de pouvoir générer une représentation graphique sous la forme de camembert de la répartition des unigènes (ESTs singleton et contigues) en grande fonction. Ceci peut être implémenté assez facilement à l'aide du logiciel R qui génère déjà ce type de graphique sous format jpeg par exemple.

3) Au niveau des différents interfaces de consultation (Sequence export, Blast abstract, search by EST/contig name), la recherche peut être faite :

– sur toutes les séquences : ESTs (singleton ou non), Contigue
– sur les ESTs (singleton ou non)
– sur les Contigues

Il est souhaité qu'on puisse faire la recherche sur :

– sur toutes les séquences : ESTs (singleton ou non), Contigue
– sur les singletons uniquement
– sur les Contigues uniquement

4) Au niveau de la page « Search by keyword », on recherche toutes les séquences (ESTs et/ou contigues) qui ont le mots clé recherchés au niveau de leur description blast. Il faudrait pouvoir effectuer cette recherche au niveau de la description blast de la séquence et au niveau de la grande fonction auquelle est associée cette séquence.

5) Amélioration de l'interface d'administration. A l'heure actuelle, seul l'administrateur (le bio-informaticien) peut créer les projets et les comptes utilisateurs. Il faudrait mettre en place un acteur gestionnaire de données qui puissent créer et modifier uniquement ses projets. Il pourrait également gérer les comptes utilisateur au niveau de ses projets. Cette amélioration est à développer en fin de projet si le temps le permet.

Livrable 3 (souhaité en Septembre/Octobre) : développer un nouveau module d'analyse à ESTdb dédié à l'annotation automatique des séquences. Le module "Annotation Automatique" doit permettre de classer automatiquement les séquences ESTs annotées par le logiciel Blast en grande fonction en se basant sur une ontologie du domaine, la Gene Ontology.

Plusieurs outils ont déjà été testés séparemment mais jamais comparés :

- Goblet : http://goblet.molgen.mpg.de/cgi-bin/goblet/webapp-goblet.cgi
- Scripts perls annotator.pl and go_fish_source.pl disponibles à l'adresse suivante : Canadian Bioinformatics Help Desk (http://www.gchelpdesk.ualberta.ca)

- Annotator.pl reads multiple sequence files in FASTA format from a file and submits each to local BLAST. The complete BLAST results are written to a file, and the best match is sent as an Entrez query to NCBI. The returned GenBank file is parsed to obtain the title of the hit, the hit's accession number, and the name of the organism that the hit was obtained from. The query sequence is written to a file in fasta format, with its title modified so that it contains information about the hit sequence, the BLAST e-value, and the BLAST bit score. Four files are created, one containing a log of the messages produced by the program, one containing the modified query titles with the query sequences, one containing just the modified query titles, and one containing the complete BLAST results obtained for each query sequence.

- Go_fish_source.pl assigns GO numbers and descriptions for blast results generated by annotator.pl. The script needs the following files to run:

  - Gene association files for uniprot (SwissProt/trembl), pdb, and Genbank: ftp://ftp.geneontology.org/pub/go/gene-associations
  - The GO DAG flat files (component, process, and function): ftp://ftp.geneontology.org/pub/go/ontology-archive
  - The blast results file from annotator.pl

To run this script, here is an example of what one would type from command line: go_fish_source.pl -i complete_blast_results123.txt -o complete_blast_results123.txt.results -e 10.

The script reads in the blast table output generated by annotator.pl and builds a hash table with the query sequence name as the key and an array of the unique gi numbers found from the blast results table as the value. A series of accession number:GO number indexes are then generated (GENBANK, SwissProt/Trembl, and PDB). The script then looks for GO numbers for each gi number and associated accession number from the blast results hash, and assigns to each gi number an array of GO numbers, if they exist in the acc num: GO number indexes. The script also provides a description of each GO number. The print out lists the query sequence, its gi numbers, and the GO number/descriptions for each gi number. A second list is printed of all the query sequences and gi numbers for which no GO number could be found. The total number of gi's with found GOs and the total number of gi numbers with no GO numbers are tabulated and printed out at the end.

D'autres logiciels existent mais n'ont pas été testé (ex : GOAnna http://www.geneontology.org/GO.tools.annotation.shtml). Il faudrait effectuer une veille technologique pour avoir une liste d'outils réalisant cette analyse, les tester et intégrer le meilleur au pipeline.

Livrable 4 : Développer un nouveau module d'analyse à ESTdb dédié)à la réalisation d'analyse de génomique comparative. Le module "Génomique comparative" permettra de réaliser des comparaisons entre les génomes des diverses plantes étudiées et/ou avec les génomes des plantes modèles et/ou avec des baqnues d'ESTs réalisées sur d'autres espèces végétales. Ces analyses optimiseront l'identification de séquences orthologues entre différentes espèces et donc l'annotation des gènes identifiés. Ces comparaisons permettront de rechercher les relations existantes entre les gènes de différentes espèces (synténie) ainsi que les relations de ces gènes au sein d'un même génome. (cf. pour information publication « Nucleic Acids Res. 2002 June 1; 30(11): 2316–2328. - « Synteny between *Arabidopsis thaliana* and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project » - Jérôme Salse, Benoit Piégu, Richard Cooke,[a] and Michel Delseny »)

Annexe

Annexe 1 : lieux d'exécution des marchés, espace mis à disposition

La prestation aura lieu sur le site de l'IRD, 911 avenue d'Agropolis dans les locaux de l'unité de recherche. Les équipements qui seront mis à la disposition du prestataire appartiennent à l'unité.

Annexe 2 : liste des logiciels et systèmes de l'unité
Architecture informatique de la plate-forme bio-informatique du centre IRD de Montpellier :

- deux serveurs de production "DellTM PowerEdgeTM 6650" (4 processeurs Xeon - 8 GB de RAM)
- une baie de disque "Dell/EMC CX300" (1 To de capacité de stockage)
- un serveur de fichier « Dell PowerEdge 1750"
- un serveur de développement (2 processeurs)

Environnement logiciel :

- outils classiques de bio-informatique : blast, sim4, clustalw, package EMBOSS...
- serveur web Apache 2.0.53
- langage de programmation : PHP 5.03, perl/bioperl
- SGBD : Mysql 3.2x.
- Environnement linux : RedHat AS 3.0

Annexe 4 : Liste des publications

Fernandez D., Santos P., Agostini C., Bon M.-C., Petitot A.-S., Silva M. C., Guerra-Guimarães L., Ribeiro A., Argout X. and Nicole M. 2004. Coffee (Coffea arabica L.) genes early expressed during infection by the rust fungus (Hemileia vastatrix). Molecular Plant Pathology ,5, 527-536.

Hocher V., Auguy F., Argout X., Laplaze L., Franche C., and Bogusz D. Expressed sequence tag analysis in Casuarina glauca actinorhizal nodule and root. New Phytologist. 2006 169:681-688.

Jouannic, S., Argout, X., Lechauve, F., Fizames, C., Borgel, A., Morcillo, F., Aberlenc-Bertossi, F., Duval, Y., and Tregear, J. (2005). Analysis of expressed sequence tags from oil palm (Elaeis guineensis). FEBS Letters 579, 2709-2714.

Kongsawadworakul P., Sookmark U., Nandris D. and H. Chrestin (2005) - Cyanide metabolism and molecular approach of rubber trunk phloem necrosis: Present and prospects (oral communication). In: Proc. Int. Hevea workshop on tapping panel dryness. Kerala, India, November 2005.

Poncet, V., Rondeau, M., Tranchant, C., Cayrel, A., Hamon, S., de Kochko, A., Hamon, P. (2006). SSR mining in coffee tree est databases: potential use of EST-SSRs as marker across Coffea genus. Mol. Genet. Geno. 276, no. 5, pp. 436-449.

Annexe 5 : Descriptif des projets scientifiques

Programme 1 : Symbioses actinorhiziennes (Equipe Rhizogenèse IRD, UMR DIA-PC ; Equipe Ecologie microbienne, CNRS/Univ Lyon 1, UMR 5557)

Les travaux développés par l'équipe IRD Rhizogenèse symbiotique ont pour objectif de comprendre les mécanismes moléculaires et cellulaires qui aboutissent à la mise en place et au développement des racines symbiotiques des arbres tropicaux de la famille des Casuarinacées (Filao). Les Casuarinacées peuvent former des nodules racinaires en symbiose avec une bactérie filamenteuse du sol, Frankia. Les Casuarinacées appartiennent au groupe des plantes actinorhiziennes qui représente après les Légumineuses le deuxième groupe de plantes fixatrices d'azote. Les arbres tropicaux de la famille des Casuarinacées jouent un rôle environnemental essentiel, notamment pour les pays du Sud. Ces plantes possèdent une croissance rapide, sont bien adaptées à la sécheresse et sont capables de coloniser des sols pauvres.
En 2002, une étude comparative du transcriptome des racines et des nodules de C. glauca a été entreprise dans le cadre du GENOPOLE Montpelliérain et le séquençage de transcrits à partir de deux banques d'ADNc de Casuarina (racine et nodules) a permis l'obtention d'environ 3000 séquences. L'analyse bio-informatique sur la plateforme IRD a permis la validation des séquences et la création de la première base de données génomique pour les plantes actinorhiziennes.

Nous développons actuellement un projet visant à comparer plusieurs espèces de plantes actinorhiziennes ayant des caractéristiques différentes en terme de mode d'infection et de développement

nodulaire. L'obtention en 2007 d'un projet Genoscope en association avec l'équipe de P. Normand (CNRS/Université Lyon I), va nous permettre le séquençage de 50 000 ESTs à partir de deux espèces actinorhiziennes (25 000 séquences pour C. glauca et 25 000 pour l'aulne, Alnus glutinosa) dont le traitement bioinformatique sera réalisé à l'IRD. Une analyse comparative des séquences produites devrait permettre l'identification in silico de gènes communs aux espèces actinorhiziennes activés lors de la mise en place de la symbiose. Par ailleurs, des études phylogénétiques récentes suggèrent une origine commune pour les différents types de symbioses fixatrices d'azote. La comparaison des séquences obtenues à celles disponibles pour les légumineuses devrait permettre de dégager des mécanismes moléculaires communs aux deux types de symbioses et ainsi de mieux comprendre les facteurs impliqués dans la mise en place des nodules racinaires fixateurs d'azote. Ce projet permettra la mise en place d'un réseau international qui contribuera au développement des ressources génomiques des plantes actinorhiziennes.

La réalisation de ce programme de recherche est conditionné par l'utilisation de la plateforme bio-informatique IRD qui nécessite (1) des optimisations pour permettre à nos partenaires un accès aux données génomiques et (2) l'ajout de nouvelles fonctions pour la réalisation des analyses de génomique comparative.

Programme 2 : Coffea (Equipe génomique et qualit é du café, IRD , UMR DIA-PC)

L'équipe génomique et qualité du café s'engage dans un programme en génomique comparative au sein des Rubiaceae et entre Rubiaceae et Solanaceae, sans pour autant négliger la comparaison avec Arabidopsis.

On dispose de plusieurs milliers de séquences EST café produits par notre laboratoire ou par d'autres membres du réseau international génomique caféier (ICGN). Dans le cadre du réseau international RubiComp (Rubiaceae comparative) soutenu par l'IRD dans sa fonction d'Agence, il est prévu dans un avenir très proche de produire plusieurs milliers d'EST à partir de différents tissus et de banques soustraites de Psychotria (Rubiaceae). La constitution d'une nouvelle banque BAC caféier est en projet et conduira dans un premier temps, au séquençage des extrémités des séquences BAC. Enfin, de très nombreuses données en génomique sous forme d'EST, de séquences de BAC et de séquençage de génome concernant la famille des Solanaceae (essentiellement la tomate) sont déjà disponibles.

Dans ce projet, nous nous intéressons à la comparaison de séquences et à l'identification de séquences orthologues intra et inter familles. Dans ce but, l'annotation homogène et cohérente des différentes banques permettra des recherches simplifiées de nouvelles séquences et constituera un système de référence. Les travaux en génomique comparative via la cartographie comparée (macro-synténie) entrepris dans notre équipe pourront être affinés à partir de l'analyse de séquences de BAC (micro-synténie).

Programme 3 : Café / cotonnier (Equipe Résistances IRD, UMR 186 RPB)

Nos objectifs sont d'identifier et de comprendre les mécanismes cellulaires, moléculaires et génétiques mis en jeu dans la résistance des plantes aux parasites. Plus précisément, nos recherches se focalisent d'une part, sur l'identification et la caractérisation fonctionnelle de gènes impliqués dans la résistance et l'activation des réactions de défense, et, d'autre part, sur l'exploration de la diversité des mécanismes de résistance associés à différentes interactions plante/parasite.

Nos modèles d'étude sont:

- le caféier (Coffea arabica) attaqué par le champignon Hemileia vastatrix, et les nématodes du genre Meloidogyne,
- le cotonnier (Gossypium hirsutum) infecté par la bactérie Xanthomonas campestris pv malvacearum.

Les activités développées font appel à des approches de génomique fonctionnelle, les approches transcriptomiques étant privilégiées. Chez le caféier, nous avons développé des banques d'ADNc soustractives pour établir un catalogue des gènes exprimés lors des réponses de résistance du caféier aux parasites et plusieurs gènes spécifiquement exprimés dans la résistance ont été clonés. Cependant, environ 35% des ESTs obtenues dans le cadre de l'interaction du caféier à M. exigua n'ont pu être annotées, faute de similarité avec des séquences connues, et pourraient représenter des séquences spécifiques des interactions plante/nématodes. D'autres banques d'ADNc sont en cours de construction avec nos partenaires Brésiliens (Embrapa) associés à ce projet, et nécessiteront l'utilisation d'outils bioinformatiques automatisés pour l'annotation des séquences. Chez le cotonnier, des approches physiologiques ont permis d'identifier plusieurs enzymes essentielles intervenant dans les voies de signalisation de la résistance (lipoxygénase, peroxydase, lipase) et les gènes correspondants sont en cours d'analyse fonctionnelle. Les recherches s'orientent vers la caractérisation de facteurs de transcription de type AP2 impliqués dans la voie de signalisation dépendante du jasmonate.

Pour les deux plantes, plusieurs milliers d'ESTs sont maintenant disponibles dans GenBank, mais ne représentent pas encore l'intégralité du génome transcrit. L'apport de la génomique comparative est donc essentiel à l'identification de nouveaux gènes, et à la caractérisation de leur fonction. Ainsi, par exemple, comme déjà précisé dans le programme 2, l'intégration des ressources génomiques de la famille des Solanaceae, proche de celle des Rubiaceae dont fait partie le caféier, permettra sans aucun doute d'identifier

des orthologues de gènes clés de la résistance des plantes aux parasites. En particulier, un gène de résistance aux nématodes a été cloné chez la tomate, alors qu'aucun gène n'est encore connu chez A.thaliana, faute de résistance aux nématodes chez cette espèce modèle. Par contre, on pourra s'appuyer sur les connaissances de la famille AP2 chez A. thaliana pour isoler leurs orthologues chez le cotonnier. Les ressources bioinformatiques qui seront développées à l'IRD faciliteront ces recherches et l'annotation des nouvelles séquences.

Programme 4 : Palmier à Huile (Equipe Arécacées, IRD , UR 192 Palmiers)

Le palmier à huile (famille Arecaceae, ordre Arecales) est une monocotylédone pérenne cultivée en zone inter-tropicale qui constitue, depuis 2004, la première source d'huile végétale dans le monde. De par sa grande productivité, cette plante est un enjeu clé pour le développement de l'agriculture durable dans de nombreux pays tropicaux mais également dans l'approvisionnement de biocarburants sources d'énérgies renouvelables au niveau mondial.

L'équipe Arécacées s'intéresse à différents aspects de la biologie du développement de cette plante dont des connaissances approfondies sont nécessaires pour pouvoir mettre à la disposition des planteurs un matériel végétal performant. Plus particulièrement, nous nous intéressons à la floraison (détermination de la structure florale, anomalies homéotiques de type épigénétique), à la fructification et à la formation de l'embryon (embryogenèses zygotique et somatique).

Afin d'étudier les processus de régulation sous-jacents à ces différents aspects du développement reproducteur, nous poursuivons, depuis plusieurs années, une approche de type transcriptomique. Ceci implique la constitution d'une collection d'étiquettes de séquence d'ADNc (EST) et l'utilisation des clones correspondants pour effectuer des expériences de type macroarray. A l'heure actuelle, la collection non redondante de séquences EST s'élève à plus de 6 000 séquences et continue de grandir. La collection de clones EST provient de plusieurs organes différents de la plante (inflorescence, pousses feuillées, embryons somatique et zygotique…).

Le palmier à huile étant relativement éloigné, d'un point de vue phylogénétique, des plantes modèles, l'annotation d'EST est plus compliquée que pour d'autres espèces. Néanmoins, l'identification de gènes rthologues chez le riz porte un grand intérêt car une séquence génomique complète est déjà disponible pour cette espèce et un nombre important et croissant de données fonctionnelles (mutants d'insertion, profils d'expression, lignées « enhancer trap »…) est disponible.

Programme 5 : Hevea (Equipe Hévéa, IRD-Mahidol University, UR060/Clifa)

Hevea brasiliensis, est la seule espèce végétale cultivée (zone tropicale humide) pour la production de latex, duquel est tiré le caoutchouc naturel.

Le programme « Recherche de marqueurs moléculaires du stress et de gènes candidats liés à la production du latex chez Hevea brasiliensis », menée par l'équipe Franco– Thaïe (IRD-Mahidol University) est basé sur l'analyse de l'expression différentielle de gènes dans la latex et le phloème (écorce interne) d'Hevea. L'étude porte sur des arbres de clones à haut et bas potentiel de production, soumis ou non à stress abiotiques (anthropiques ou environnementaux) conduisant à une surproduction transitoire (agents stimulants), ou au contraire à la cessation définitive de la production du latex (syndrome des « encoches sèches » ou de la « nécrose du phloème »).

L'étude est basée sur la construction et l'analyse de banques soustraites (SSH) d'ADNc de latex ou d'écorce interne des différents phénotypes étudiés. Six banques SSH ont déjà été construites en 2005 puis fin 2006, desquelles en tout environ 7.000 EST ont été séquencés. Quatre nouvelles banques SSH seront élaborées en 2007, avec un séquençage prévu d'environ 4600 nouvelles EST. D'autres programmes de séquençage à partir de nos banques d'ADNc pleine longueur, sont prévus dans un futur proche.

L'analyse bioinformatique des ces banques d'EST, au moyen du pipeline EST-DB de l'IRD-Montpellier, permettra le tri d'unigènes et l'élaboration, dans un premier temps, de filtres macroarray, puis à terme de microarrays. Ces futurs outils serviront au diagnostic pour l'optimisation de l'exploitation en plantation, et pour la sélection précoce de nouveaux clones performants, dans le cadre des programmes d'amélioration de l'hévéa, mis en œuvre au sein des instituts spécialisés des différents pays de la zone tropical humide, producteurs de caoutchouc naturel.

Ce programme nécessite l'utilisation de gros moyens de calcul et de fortes compétences en matière de bioinformatique et statistique. Dans le cadre de ce programme de recherche formation sur l'hévéa, deux chercheurs Thais, l'un de l'Université de Mahidol et l'autre de l'Institut BIOTEC (Bangkok), suivent une formation (2006-2008) en Mastère de Bioinformatique à l'UM2, avec stage pratique à l'IRD-montpellier sous la responsabilité de Christine Tranchant. Outre l'aide qu'il procurera au programme de recherche « Hevea », ce programme de formation, cofinancé par le MAE et le Ministère des Universités Thaïlandais, a pour but, à terme, d'initier un réseau d'agro-bioinformatique Thaïlandais, en coopération avec l'équipe de bio-

informatique/GeneTrop de l'IRD-Montpellier. A cet effet, la plateforme bio-informatique est d'ores et déjà accessible par nos partenaires thailandais via le web, notamment l'application ESTdb.